SERIAL AVERAGING IN THE CONSTRUCTION AND VALIDATION OF PERFORMANCE TESTS

Marshall B. Jones

July 1991

Final Report

This report was prepared under the Navy Manpower, Personnel, and Training R&D Program of the Office of the Chief of Naval Research under Contract N00014-90-J-1994.

The Pennsylvania State University College of Medicine

Hershey, Pennsylvania

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S): TR-91-01	5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION: The Penna. State Univ. Coll. of Med. Dept. of Behavioral Science	6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: Cognitive Science Program, Office of Naval Research (Code 1142CS), 800 North Quincy Stree
6c. ADDRESS (City, State, and ZIP Code): 500 University Drive Hershey, PA 17033		7b. ADDRESS (City, State, and ZIP Code): Arlington, VA 22217-5000
8a. NAME OF FUNDING/SPONSORING ORGANIZATION	8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: N00014-90-J-1994
8c. ADDRESS (City, State, and ZIP Code)		10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
0602233N	RM33M20		

11. TITLE (Include Security Classification): Serial Averaging in Construction and Validation of Performance Tests (Unclassified)

12. PERSONAL AUTHOR(S): Jones, Marshall Bush

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM 1 Jun '90 TO 31 May '91	9 July 1991	67

16. SUPPLEMENTARY NOTATION: Supported by the Office of the Chief of Naval Research Manpower, Personnel, and Training R&D Program

17. COSATI CODES		
FIELD	GROUP	SUB-GROUP
05	08	
	02	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Performance testing, serial averaging, reliability, temporal stability, predictive validity

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

The advent of the microcomputer has led to a renaissance in performance testing, that is, tests which sample what a person can do (remember, track, aim, detect, recognize, etc.) rather than what he or she knows. Psychometric theory, however, is based on knowledge tests. The unit of analysis is an item and the order of administering the items is arbitrary. In performance testing the unit of analysis is a trial and order of administration is not only nonarbitrary but often the only thing that distinguishes one trial from another. In a knowledge test it is not unreasonable to suppose that mean performance and interitem correlations are independent of order of administration. In a performance test it is. Typically, performance improves with practice and intertrial correlations tend toward a definite pattern as a function of order.

The consequences of these differences for theory are drastic. In performance testing, both reliability and temporal stability frequently encounter optima as a test is lengthened. Hence, low reliability or stability may not be corrigible by increasing test length. Further,

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL: Dr. Charles Davis	22b. TELEPHONE (Include Area Code): (703) 696-4096 / 22c. OFFICE SYMBOL: ONR 1142CS

DD FORM 1473, 84 MAR — 83 APR edition may be used until exhausted. All other editions are obsolete. — SECURITY CLASSIFICATION OF THIS PAGE

scoring all trials administered (the usual practice) may not yield the best obtainable predictive validity. Scoring only a subset of consecutive trials (early, middle, or late) frequently yields appreciably higher predictive validities than the conventional practice.

"Subset analysis" serves the same ends in performance-test theory as item analysis does in conventional psychometrics. Both kinds of analysis concern the selection of some materials for inclusion in a test and others for exclusion, either in original development or in subsequent revision. The difference is that item analysis focuses on individual items and subset analysis on subsets of ordered trials.

Serial averaging and its applications (reliability and stability optima, optimal scoring for predictive validity, and subset analysis) are explained and illustrated. Results obtained using the Project-A computer-administered tests serve as the database.

ABSTRACT

The advent of the microcomputer has led to a renaissance in performance testing, that is, tests which sample what a person can do (remember, track, aim, detect, recognize, etc.) rather than what he or she knows. Psychometric theory, however, is based on knowledge tests. The unit of analysis is an item and the order of administering the items is arbitrary. In performance testing the unit of analysis is a trial and order of administration is not only nonarbitrary but often the only thing that distinguishes one trial from another. In a knowledge test it is not unreasonable to suppose that mean performance and interitem correlations are independent of order of administration. In a performance test it is. Typically, performance improves with practice and intertrial correlations tend toward a definite pattern as a function of order.

The consequences of these differences for theory are drastic. In performance testing, both reliability and temporal stability frequently encounter optima as a test is lengthened. Hence, low reliability or stability may not be corrigible by increasing test length. Further, scoring all trials administered (the usual practice) may not yield the best obtainable predictive validity. Scoring only a subset of consecutive trials (early, middle, or late) frequently yields appreciably higher predictive validities than the conventional practice.

"Subset analysis" serves the same ends in performance-test theory as item analysis does in conventional psychometrics. Both kinds of analysis concern the selection of some materials for inclusion in a test and others for exclusion, either in original development or in subsequent revision. The

difference is that item analysis focuses on individual items and subset analysis on subsets of ordered trials.

Serial averaging and its applications (reliability and stability optima, optimal scoring for predictive validity, and subset analysis) are explained and illustrated. Results obtained using the Project-A computer-administered tests serve as the database.

INTRODUCTION

The theoretical problem of performance testing

The distinction between knowledge and performance testing turns on what one is trying to measure. A knowledge test samples what a subject knows, a performance test what he or she can do. Plainly, this distinction is not absolute. A mathematics test, for example, may involve not only what a subject knows but also what he or she can do with that knowledge. A memory task may be facilitated if a subject has seen an unusual symbol before and knows what it is, say, a Greek omega. Nevertheless, most tests fall lopsidedly into one category or the other.

In a knowledge test the subject does not usually know whether the answer that he or she has given is right or wrong. As a result practice effects are limited to auxiliary aspects of the test (test-taking skills) and, while they exist, are not large (Messick & Jungblut, 1981; Wing, 1980). In a performance test, however, it is usually not possible to prevent the subject from obtaining some idea as to how well or poorly he or she is doing. As a consequence, subjects tend to do better on a test the more times it is administered to them (Bittner et al, 1983; Kennedy et al, 1981). In effect, each test administration becomes a trial of practice.

Psychometric theory is based on knowledge tests. The unit of analysis is an item and the order of administering the items is arbitrary. In performance testing, however, the unit of analysis is a trial and order of administration is not only nonarbitrary but often the only thing that distinguishes one trial from another. In a knowledge test it is not unreasonable to suppose that mean performance and interitem correlations are independent of order of administration. In a performance test it is.

Typically, performance improves with practice, often within a session but almost always from test to retest; variances follow the means; and intertrial correlations tend toward a definite pattern as a function of order, the superdiagonal form (Jones, 1962).

The practical problem of performance testing

During the Second World War performance testing based on electromechanical apparatus (rotary pursuit, complex coordination, two-hand tracking, and the like) was widely and successfully used in military selection, especially for pilot training (Melton, 1947). The equipment, however, was heavy, bulky, difficult to maintain, and more difficult to replace. By the late 1950s all three military services had abandoned performance testing in favor of paper-and-pencil tests exclusively. Then in the late 1970s the advent of microcomputers reopened the possibility of performance testing, this time with equipment that occupied little space, did not break down frequently, and was easily replaced when it did. At the same time, experimental psychology was undergoing a revolution of its own, as the discipline's central focus shifted from learning theory to cognition and information-processing. The joint effect of these two developments was a new generation of cognitively oriented, microcomputer-based performance tests (Englund, Reeves, Shinglebecker, Thorne, Wilson, & Hegge, 1987; Kennedy, Baltzley, Wilkes, & Koontz, 1989; Kyllonen & Christal, 1989).

Unfortunately, all has not been clear sailing for this new generation of performance tests. The most serious problem has been that many tests have low reliabilities (Kyllonen, 1985). Predictive validities against real-world criteria are still sparse, but it seems likely that oftentimes they will also be low. An appropriate response to these difficulties involves more than

making and trying out new tests. What is needed is a theory of performance tests, that is, an approach to test construction and validation that recognizes and capitalizes upon the distinctive properties of performance tests.

Plan of the paper

The present paper develops such a theory. Its empirical base is provided by the ten computer-administered tests in Project A (Eaton, Hansen, & Shields, 1987; Peterson, Hough, Dunnette, Rosse, & Wing, 1990). Results will be presented under seven headings: comparisons with Army data, practice effects, reliability, temporal stability, sample variations, predictive validity, and subset analysis.

TASKS, SUBJECTS, PROCEDURES

The Project-A tests

Project A is a large, multi-year effort to improve the Armed Forces Vocational Aptitude Battery (Eaton, Hanser, & Shields, 1985; Peterson, 1987). Included in this effort are ten newly developed, computer-administered performance tests. Brief descriptions of the ten tests are given below. The tests are administered in the order described. Table 1 shows the number of trials a subject receives on each test and, approximately, the total length of time each test requires.

Simple Reaction Time. The subject is instructed to place his or her hands in the ready position. When the word YELLOW appears in a display box, the subject strikes the yellow key on the test panel as quickly as he or she can. The dependent measure is average time to respond.

Choice Reaction Time. This test is much the same as Simple Reaction Time. The major difference is that the stimulus in the display box is BLUE or

WHITE (rather than YELLOW), and the subject is instructed to strike the corresponding blue or white key on the test panel. The dependent measure is average time to respond on trials in which the subject makes the correct response.

Short-Term Memory. A stimulus set, consisting of 1, 3, or 5 letters or symbols, is presented on the display screen. Following a delay period, the set disappears. When the probe stimulus appears, the subject must decide whether or not it was part of the stimulus set. The dependent measure is average time to respond on trials in which the subject makes the correct response.

Target Tracking 1. This is a pursuit tracking test. The subject's task is to keep a crosshair centered within a box that moves along a path consisting exclusively of vertical and horizontal lines. The dependent measure is the average distance from the crosshair to the center of the target box.

Perceptual Speed and Accuracy. This test measures a subject's ability to compare rapidly two stimuli presented simultaneously and determine whether they are the same or different. The stimuli may contain 2, 5, or 9 characters and the characters may be letters, numbers, or other symbols. The dependent measure is average time to respond on trials where the subject's response is correct.

Target Tracking 2. This test is the same as Target Tracking 1, except that the subject uses two sliding resistors instead of a joystick to control the crosshair. The dependent measure is the same as in Target Tracking 1.

Number Memory. The subject is presented with a number on the computer screen. When the subject presses a button, the number disappears and another

number appears along with an operation term (e.g., "Add 9" "Multiply by 3"). When the subject presses a button, another number and operation term are presented. This procedure continues until finally a solution to the problem is presented. The subject must then indicate whether the solution presented is correct or incorrect. The dependent measure is total time to respond on trials in which the subject correctly identifies the solution presented as correct or incorrect.

Cannon Shoot. The subject's task is to fire a shell from a stationary cannon so that it hits a target moving across the cannon's line of fire. The dependent measure is a deviation score indicating the difference between time of fire and optimal fire time (for example, direct hit yields a deviation score of zero).

Target Identification. The subject is presented with a target and three stimulus objects. The objects are pictures of tanks, planes, or helicopters. The target is the same as one of the three stimulus objects but rotated or reduced in size. The subject must determine which of the three stimulus objects is the same as the target object. The dependent measure is average time to respond on trials in which the subject makes the correct response.

Target Shoot. The subject's task is to move a crosshair over a moving target and then press a button to fire. The dependent measure is distance from the crosshair to the center of the target when the subject fires.

The criterion task

In addition to the Project-A tests, each subject was administered a criterion task. This task was Anti-Aircraft, game #1 in the Atari Air-Sea Battle cartridge (CX-2624). In this game the subject controls a gun placed two thirds of the way from left to right at the bottom of the television

screen. Four different kinds of aircraft traverse the screen above the gun, in different numbers, at different speeds and altitudes, and from left to right or vice versa. The purpose of the game is to shoot down as many aircraft as possible in a 2-min-and-16-sec game. The control devices are a joystick for positioning the gun and a button for firing the missile. The missile itself was the smaller of two possible sizes (difficulty position "A"). The dependent measure is number of aircraft shot down per game.

Anti-Aircraft is a complex psychomotor skill with a high ceiling. No subject comes close to reaching the maximal possible performance with the amount of testing given.

Subjects and procedures

The subjects were two independent samples of undergraduate students at central Pennsylvania colleges. Both samples numbered 102 subjects, 50 men and 52 women in Sample A and 49 men and 53 women in Sample B. The two samples were collected at the same colleges two years apart, Sample A in 1988-89 and Sample B in 1990-91. Design and procedures were identical in the two samples.

Each subject was administered the Project-A tests at the start of the fall semester (September, October) and then again four months later at the start of the spring semester (January, February). The Project-A tests were taken in a single sitting that lasted between 45 and 75 mins, depending on how quickly the subject responded to the tests and the instructions that preceded them. The entire administration, both test and retest, instructions as well as the tests themselves, was computer-controlled.

In the fall, following the Project-A tests, each subject was administered five sessions of Anti-Aircraft, each session consisting of seven

games or a little less than 16 mins of playing time. All five sessions were completed within a ten-day period, with no more than two sessions taking place on a given day. In the spring semester, again following the Project-A tests, each subject was given three sessions of Anti-Aircraft with the same number of games per session and the same conditions as to distribution as in acquisition.

COMPARISONS WITH ARMY DATA

Table 2 compares the present results (Sample A) with those collected by Peterson, Hough, Dunnette, Rosse, Houston, Toquam, and Wing (1990) in overlapping samples of Army enlisted people ranging in number from 8,892 to 9,269, depending on the test. The tests were scored the same way at Hershey as in the Army, that is, a subject's score on any given test is the average of his or her score on all trials administered.

The college students perform better on all tests, but some of the differences are sizable whereas others are trivial. The largest differences are for the two memory tests, in both cases half a standard deviation (s) or more. The next largest differences are for the two "perceptual" tests (Perceptual Speed & Accuracy and Target Identification), approximately .4s. The differences for Choice Reaction and the two tracking tests are approximately .33s, while those for Simple Reaction and the two aiming tests (Cannon Shoot and Target Shoot) are less than .2s. These differences are broadly what one would expect; the more "cognitive" a test is the larger the difference in favor of the students tends to be.

Variabilities were greater in the Army than in the Hershey data, except for Target Tracking 2, but not greatly so, except for Simple Reaction. The variance of Simple Reaction is nine times as large among the enlisted people

as among the students. Simple Reaction is the first test in the battery, and there may have been some confusion among the Army subjects as to what they were supposed to do. If so, it would explain the high variability of Simple Reaction in the Army data.

The column headed "Reliability" contains, for the Army data, odd-even correlations corrected for test length by the Spearman-Brown formula and, for the Hershey data, Spearman-Brown projections from the average correlation involving all trials. Thus, both figures make use of all trials administered and both use the Spearman-Brown formula. The correspondence between the two sets of figures is startingly close.

The column headed "Temporal Stability" contains two-week test-retest correlations for the Army data and four-month test-retest correlations for the Hershey data. Temporal stability was better at Hershey than in the Army for all tests except Target Identification and may have been better even for Target Identification, given that the retest interval was eight times as long at Hershey as in the Army.

There are at least four subject or procedural differences that may have contributed to the better stability at Hershey. First, of course, was the difference in population: college students versus enlisted people. Second, the sex ratio at Hershey was essentially 50-50, whereas males predominated in the Army sample. Third, the Hershey tests were administered by a single, very experienced person, whereas the Army data were collected at many places by many people, some of them not experienced test administrators. Fourth, the Hershey subjects were tested one or two at a time, whereas the Army subjects were tested in batches of as many as two or three dozen at a time.

In general, however, the differences between the two arrays of stability results are not large. The low stability for Simple Reaction in the Army data is probably related to that test's high variability. No obvious explanation exists for the low stability of Target Shoot in the Army sample, except perhaps that Target Shoot is the last test in the battery.

PRACTICE EFFECTS

Practice effects occur regularly in performance tests and in several different forms. Figure 1 illustrates the most common effect. The figure presents mean results for the two tracking tests in Sample A at test and retest. The means in Figure 1 are not means of individual trials. The score for a given subject at trial i is the average of his or her scores up to and including that trial, what I will call a "forward average." The means in Figure 1 are means of forward averages. Forward averaging is done separately within test and retest sessions.

The first trial in both tests happens to be easy. Hence, the mean error score is small initially for both tests. Thereafter, however, mean performance shows little change. There is, however, a marked and highly significant fall-off ($p<.001$) from test to retest for both tasks. If one compares the final points at test and retest, that is, the averages of all trials administered, the subjects perform better at retest than at test on all ten tests in the Project-A battery and in five of the ten tests the difference is significant at the .01 level or better, taking both samples into account.

A second effect, also evident in Figure 1, is that the difference between test and retest is generally larger in the early trials than later on. In Target Tracking 1, for example, the difference for Trial 1 is .24 log

units and for the first two trials .20. At the end of practice, the averages for all 18 trials, the difference is .14 log units. These figures are representative. In a typical case the subjects' initial performance is markedly better at retest than at test, but the difference narrows with continued testing. A plausible interpretation is that the subjects learn how to respond from their first exposure and this learning gets them off to a better start at retest but does not help them as much or, perhaps, at all after the first few trials.

Trials on Memory fall into three subsets, according to size of the stimulus set. In 12 of the trials the set consists of a single stimulus, in 12 others of three stimuli, and in the remaining 12 trials of five stimuli. In all 36 trials the subject is subsequently presented with a probe stimulus and asked to indicate whether or not it was included in the stimulus set. Figure 2 presents mean results for the three subsets of Memory at test and retest in Sample A. Again, each individual's score is a forward average, that is, the average of his or her scores up to the trial indicated.

Mean performance for all trials improves from test to retest for all three subsets but the difference is significant only for Subset 3. The difference for Trial 1, however, is significant at the .01 level for Subsets 3 and 5 and at the .05 level for Subset 1. These effects have already been noted as typical of performance tests in general. Memory, however, shows two additional effects that are seen only in some tests.

The curve for Subset 3 decreases significantly ($p<.01$) from Trial 4 to Trial 12 at test, and this decrease recurs in Sample B ($p<.01$). In the case, therefore, of this subset it would appear that there is evidence for learning within the test session as well as between it and the retest session. The

curve for Subset 1 decreases nonsignificantly from Trial 4 to Trial 12, while the curve for Subset 5 increases nonsignificantly over the same span of trials.

Increasing trends are prominent at retest. The curves for Subsets 1 and 5 both increase significantly ($p<.01$) from Trial 4 to Trial 12 at retest and, again, these trends recur in Sample B ($p<.01$). These increasing trends are also practice effects, although, of course, they cannot be interpreted as evidence for learning. The most plausible interpretation is that they reflect a practice-induced fatigue or loss of concentration, which should be more prominent at retest than at test. It does appear, however, that in the most difficult subset (5) the increase from Trial 4 to Trial 12 occurs in the test session also. In Sample A the increase is not significant but in Sample B it is ($p<.01$).

Variances, in general, follow the mean. Table 3 presents means and variances for Memory at test and retest, broken down by sample and subset. Individual scores are averages of all 12 trials in a subset. As can be seen, the coefficient of variation ranges between .21 and .27 despite large differences in the means. This observation does not mean, of course, that practice has no effect on variances, only that these effects rarely, if ever, contain any information additional to that contained in the means.

In studies of skill acquisition, correlations between sessions of practice fall into a regular and highly reproducible pattern, the superdiagonal form (Jones, 1962, 1969). The gist of this pattern is that the closer together two sessions are in the practice sequence the stronger the correlation between them. Neighboring sessions correlate most strongly, while the weakest correlation is between the first and last session. How

clearly this pattern appears depends primarily on sample size and the amount of performance represented by a single data point. Where each data point is based on many minutes of performance, the superdiagonal pattern is always seen and is usually quite regular. For example, the five test sessions of Anti-Aircraft, where each session lasts 16 minutes, show it very clearly. Intertrial correlations between individual trials, each one lasting only a few seconds, are another matter. The pattern may be there but in order to show it the trials must be grouped into blocks.

Table 4 presents the intertrial correlations for Perceptual Speed and Accuracy in Sample A at test. The trials have been grouped into blocks of four trials each. Hence, a correlation between two different blocks is the average of 16 intertrial correlations, while the correlation within a block is the average of six intertrial correlations. The latter correlations appear in the main diagonal in parentheses.

Even with blocking, the superdiagonal pattern in Table 4 is somewhat irregular. Nevertheless, it is unmistakably present. The correlations in the main diagonal are, on the average, larger than those in any other diagonal. The next largest are in the superdiagonal, the next diagonal over, containing correlations between neighboring blocks. With each successive diagonal the correlations become smaller until one reaches the upper, right-hand corner, which contains the smallest correlation in the matrix, .17.

In some tests, the two tracking tests are cases in point, the superdiagonal pattern is very shallow. In others, for example, Cannon Shoot and Target Shoot, the level of correlation is very low and, therefore, individual correlations are extremely variable. Nevertheless, shallow or obscured by variability as it may be, superdiagonal pattern is a persistent

feature of performance tests. It is also a practice effect. Superdiagonal form does not necessarily reflect learning but it always reflects temporal or sequential order. Like learning or loss of concentration, it is induced by trials of practice.

Altogether, then, practice effects have been noted in five distinct forms: mean improvements from test to retest, mean improvements within a test session, mean deterioration within a test session, trends in variance with practice, and intertrial correlations tending toward superdiagonal pattern. These various effects pose many, but not necessarily insoluble problems for performance testing. One of them concerns reliability as a function of test length.

RELIABILITY

In a superdiagonal pattern the later a trial comes in the test sequence the weaker its correlation is with a given early trial. Put differently, intertrial correlations decrease along any row to the right. Table 5 presents a hypothetical superdiagonal pattern. As can be seen, the correlations decrease regularly by .05 along any row to the right. This feature of superdiagonal pattern has definite implications for reliability as a function of test length.

In conventional test theory the Spearman-Brown (S-B) formula (Gulliksen, 1950) states that the reliability of a test i units in length

$$R_i = \frac{i\,R_1}{1 + (i-1)\,R_1},$$

where R_1 is the reliability of a test of unit length. When $i \geq 2$, R_1 is taken as the average correlation among the i units, that is, $\bar{r}_i$. The first row at the bottom of Table 5 shows this average correlation--for the first two

trials, the first three, out to all seven trials. As is clear from the table, these averages decrease as one moves forward from the first to the last trial. Because the correlations decrease along any row to the right, each new trial adds to the average a column of correlations lower than those already in it; hence $\bar{r}_i$ drops a notch.

Low reliability in a knowledge test is corrigible. It may be laborious to do, but in principle one can always lengthen the test, while maintaining the same average inter-item correlation, and thereby improve its reliability. In a performance test, however, $\bar{r}_i$ may not remain the same as the test is lengthened; in most tests it decreases. The bottom row in Table 5 gives R_i as calculated by the S-B formula for $i = 1,\ldots, 7$. As $\underline{i}$ increases, $\bar{r}_i$ both decreases and is more strongly amplified by the S-B formula. The amplication, however, is negatively accelerated while, in this example, the decrease in $\bar{r}_i$ proceeds at a constant rate. The upshot is that R_i increases sharply at first, reaches a maximum (at $i = 4$), and then decreases gently. In this case, therefore, reliability would not be improved by lengthening the test. In fact, the test could be shortened to 4 trials with no loss of reliability.

The superdiagonal pattern in Table 1 is perfectly regular; that is, correlations are constant within any given diagonal and regularly decreasing between diagonals. Superdiagonal patterns are not necessarily level, however. In many psychomotor tests they have a tendency to rise with practice (Reynolds, 1952) and, where this is the case, the tendency for $\bar{r}_i$ to fall with practice may be nullified or even reversed. By the same token, however, correlational level may also fall with practice for other reasons than superdiagonal patterning. Fatigue or loss of concentration may manifest

itself in correlational levels and patterns as well as in mean performance. In the presence of fatigue or wavering attention performance tends to be fitful and erratic, which introduces novel variance not present in earlier trials of practice. The effect is to produce a drop in correlational level and, therefore, to bring about a reliability optimum earlier than it would have occurred in a perfectly regular superdiagonal pattern (but see the discussion under Comment).

Figure 3 presents reliability results for Simple Reaction. The average correlation up to trial i (solid squares) tends to decrease sharply as i goes from 2 to 10. A straight line has been fitted to these nine points and extended out to trial 25. The corresponding reliabilities (solid circles) are Spearman-Brown projections (R_i) for a test of length, i, given that a test of unit length has reliability $\bar{r}_i$. The smooth curve for R_i was obtained by applying the S-B formula to corresponding points, $\bar{r}'_i$, on the regression line. The smooth curve has also been extended to Trial 25. Such a curve reaches a maximum at

$$i^* = \frac{(1-a)-\sqrt{1-a}}{b},$$

where a and b are the intercept and slope of the regression line. In this case i* equals 19.2. It would seem, therefore, that the reliability of Simple Reaction could be improved by lengthening the test but only modestly. Roughly doubling the number of trials would increase reliability by .02 but still leave it at .897, well short of unity. More than doubling the number of trials would be counterproductive.

Figure 3 can be improved in two key respects. First, the regression line in Figure 3 was obtained by weighting the $\bar{r}_i$ equally. The $\bar{r}_i$, however,

are based on very different numbers of correlations. For example, $\bar{r}_2$ is based on only one correlation, whereas $\bar{r}_{10}$ is based on 45. It would make sense on strictly statistical grounds to weight the $\bar{r}_i$ for the number of correlations on which each one is based. It makes especially good sense when one remembers that the main purpose in fitting the regression line is to predict the course that $\bar{r}_i$ will follow beyond the administered number of trials (n). The $\bar{r}_i$ often follow a decreasing, negatively decelerated course. Therefore, the best prediction of where $\bar{r}_i$ will lie when i>n is the slope of the $\bar{r}_i$ curve, not overall, but just before the administered sequence reaches its end. Weighting the $\bar{r}_i$ for the number of correlations on which each one is based effectively approximates such a slope. The early points are heavily discounted in favor of the last few points. The resulting line in such cases is shallower than the one obtained by equal weighting of the $\bar{r}_i$. Hence, the number of trials for optimal reliability, i*, is increased (pushed further out).

The second key improvement concerns how to estimate R_1. The estimation based on $\bar{r}_i$ assumes that all trials have equal variances. If this is not so (and it never is), the appropriate estimate becomes

$$R_1 = \tilde{r}_i = \frac{\overline{cov}_i}{\overline{var}_i},$$

where $\overline{cov}_i$ and $\overline{var}_i$ are, respectively, the averages of all covariances and variances up to trial i. In effect, $\tilde{r}_i$ weights the correlations for the variances involved in them. Correlations between trials with large variances count for more than correlations between trials with small variances. This improvement has no systematic effect on i*. Sometimes it increases i* and sometimes, as in the case of Simple Reaction, it decreases i*.

Figure 4 presents the reliability results for Simple Reaction, making these two improvements. The net effect is to decrease i* to 14.8 and to reduce the optimal reliability to .874.

Table 6 presents reliability results for all ten tests, using $\tilde{r}_i$ and a weighted regression line. For two tests (Perceptual Speed & Accuracy and Target Tracking 2) the regression line has positive slope (b>0). In these two cases there is no optimal reliability short of unity. In the other eight tests slope is negative, i* finite, and optimal reliability some value less than unity. In five of the eight cases, however, i* is remote and, with the exception of Number Memory, R'_{i*}, the projected optimal reliability, is not much less than unity. In one case (Cannon Shoot), however, i*<n. That is, the number of trials for optimal reliability is less than the number administered. In such a case reliability cannot be at all improved by lengthening the test. In fact, the test could be shortened without reducing optimal reliability. In two tests, Target Shoot as well as Simple Reaction, i* lies just five trials ahead of where the administered sequence stops. In both cases little is to be gained by increasing the number of trials and optimal reliability lies well short of unity.

TEMPORAL STABILITY

In temporal stability one has two sequences of trials to consider, both test and retest, instead of just one. The procedure used in reliability, however, generalizes naturally to stability. For each individual one obtains forward averages at test (i) and at retest (j) and then calculates

the n correlations where i = j. For purposes of projection, the temporal stability coefficient is analyzed into three components:

$$_{t}s_{ij} = \left(\frac{i\tilde{\pi}_i}{1 + (i-1)\tilde{\pi}_i}\right)^{\frac{1}{2}} \frac{\overline{cov}_{ij}}{(\overline{cov}_i \overline{cov}_j)^{\frac{1}{2}}} \left(\frac{j\tilde{\pi}_j}{1 + (j-1)\tilde{\pi}_j}\right)^{\frac{1}{2}}.$$

The first and third components are the S-B expansion terms, exactly as they would appear were we calculating S-B reliability at trial i or j from the test or retest results. Taken together, these two components are the geometric mean of the test and retest reliabilities. The middle term, called the "covariance ratio," is the ratio of the average covariance between test and retest to the geometric mean of the average covariances within test and retest. This ratio is an upper bound to temporal stability, but not a correlation coefficient. The covariance ratio may exceed unity and often does. When it does, of course, it is not a least upper bound because unity is then less than it and unity is also an upper bound to temporal stability. Finally, the decomposition of temporal stability into the covariance ratio and the two expansion terms is exact. That is, if one calculates the components and multiplies them together, the result is exactly the same as calculating temporal stability directly (see Technical Note 1).

If the covariance ratio decreases as i (and j) increase, the fact is a sufficient condition for temporal stability to reach an optimum. Figure 5 presents a case in point, Choice Reaction in Sample A. A straight line has been fitted to the covariance ratios, weighting each ratio for the number of between-covariances on which it is based. For purposes of presentation, the regression lines for $\tilde{r}_i$ and $\tilde{r}_j$ have been contracted into a single curve, obtained by plotting the geometric mean of corresponding points on the two

regression lines. This one curve, it should be noted, is not in general a straight line. It should also be noted that the difference between the geometric mean of the two expansion terms and the S-B expansion of the geometric mean of the two regression lines is generally negligible. The curve for temporal stability was obtained by applying the decomposition formula for temporal stability, given above, to the fitted values for $\tilde{r}_i$, $\tilde{r}_j$, and the covariance ratios.

The empirically obtained temporal stability for Trial 27 is indicated in Figure 5 as a "false optimum." The point here is that if one had only the empirically calculated stabilities, the value at Trial 27 would be larger than any value either before or after it and might, therefore, be considered optimal. The difficulty with so identifying an empirically obtained value is that one could easily be capitalizing on chance. Any empirically obtained value contains some amount of error. The hazard, therefore, of identifying a value as optimal when, in fact, its "optimality" may be a chance upward deflection is considerable. Fitting a smooth curve for temporal stability has, of course, the merit that it allows us to project values for temporal stability beyond where the administered sequence ends. It also has the advantage of basing a reliability or stability optimum on the entire set of obtained results rather than a single data point. The point where stability reaches an optimum cannot be obtained in closed form, as can the corresponding point for reliability. Hence, stability optima must be obtained by numerical means. In the present case the effect is to push the point of optimal stability from Trial 27 out to a little past Trial 45.

Decreasing covariance ratios are a sufficient but not a necessary condition for temporal stability to reach an optimum. If $\tilde{r}_i$ and $\tilde{r}_j$ are

decreasing and especially if they are low, temporal stability may still increase to an optimum and then decrease--even if the covariance ratios are rising. Figure 6 presents stability results for Target Shoot. Intertrial correlational levels are very low, less than .10 within sessions. At these levels, when $\tilde{r}_i$ and $\tilde{r}_j$ decrease at even modest rates, the proportional drops in the corresponding expansion terms tend to be substantial. Because covariance ratios lie at a much higher level, in the present case, on the order of .90, increases in the covariance ratios tend to be small proportionally. The decomposition formula for temporal stability, however, is multiplicative. What matters are proportional changes. Hence, it may easily happen that small decreases in $\tilde{r}_i$ and $\tilde{r}_j$ are more than enough to match much larger absolute increases in the covariance ratios.

Table 7 presents stability results for all ten tests. The retest regression slopes (b_2) are negative in eight of the ten tests, just as were the test regression slopes (b_1 in Table 6). There is, moreover, a good deal of correspondence between b_1 and b_2. Simple Reaction, for example, has much the most negative slope at both test and retest. Slope for the covariance ratios is positive in eight of the ten tests. In general, the covariance ratios tend to rise as i and j increase.

These opposing trends give rise to a rather sharp dichotomy. First, stability reaches an optimum in five of the ten tests, and in five it does not; reliability optima were reached in eight of the ten tests. Second, all five of the stability optima but only three of eight reliability optima are binding. Third, the optimal stabilities are much lower than the optimal reliabilities. Stability, in short, seems either to reach an optimum early, in the present sample, prior to Trial 50, or not at all. By the same token,

the maximal stability attainable is either relatively low, in the present sample, less than .80, or unity--in one instance, indeterminate (see the footnote for Target Tracking 2).

SAMPLE VARIATIONS

Reliability and stability are sample statistics and, like any other sample statistic, subject to variation from one sample to the next. How large these variations are likely to be is best determined by deriving the sampling distribution or, failing that, by approximating it at selected points (specified by sample size, number of trials administered, correlational levels, etc.) using numerical methods. Neither of these efforts is attempted in this paper. We do, however, have two independent samples of 102 subjects each and can, therefore, obtain a crude preliminary notion of how much variation one may expect from one sample to the next.

Table 8 presents optimal-trial numbers for reliability and stability in the two samples or notes that none was found. Three of the tests in Sample A show binding reliability optima and two of these tests (Cannon Shoot and Target Shoot) show them again in Sample B. The remaining seven tests show more distant optima or none at all in both samples. Altogether, five of the tests (Target Tracking 1 and 2, Cannon Shoot, Target Shoot, and Target ID) show reasonably consistent results in the two samples. Two (Memory and Choice Reaction) are moderately discrepant, and three (Simple Reaction, Perceptual S & A, and Number Memory) sharply so.

For reasons that will be given later under Comment, temporal stability is theoretically preferable to reliability as well as practically more relevant. It also appears to be more consistent. Eight of the ten tests (all but Perceptual S & A and Target ID) are reasonably consistent.

Temporal-stability curves--or reliability curves too, for that matter--rise slowly to an optimum and decrease slowly after it. It is not surprising, therefore, that the precise location of an optimum may vary by 10, 20, or 30 or more trials from one sample to another. By the same token, however, the exact location of the optimum does not matter a great deal. The difference in stability (rather than trial numbers) is small, as a rule, not more than .01 or .02.

The optimum, for example, for Choice Reaction in Sample A is 45.7 trials, three times further out than in Sample B. Yet this discrepancy makes very little difference. Increasing the test in Sample A to 45 trials would indeed improve its stability--but by less than .01. What matters most is the existence of a stability optimum. Figure 7 presents the stability results for Choice Reaction in Sample B. The decline in the empirically obtained stability coefficients is apparent. Further, the similarity of this figure and the one for Sample A (Figure 5) is striking, despite the threefold difference in i*. Even with two samples of only 102 subjects, it seems clear that no appreciable gain in stability can be had by increasing the length of Choice Reaction. In its case, other approaches to increasing stability, for example, administering the test in two or more bouts of, say, 15 trials each, should be considered.

On the other hand, in those tests where an optimum does not exist, stability continues to increase as i and j increase, as far as can be projected, with no limit other than unity. Neither of the two tracking tests, for example, show an optimum in either sample. Their stabilities, as we know, are already high. It would appear, however, that by increasing test length they could both be made even more stable.

Altogether, stability optima tend not to exist at distant or remote trials. The optimum either does not exist or is binding. Further, most tests seem to fail consistently into one or the other grouping. The probability of as much agreement between Samples A and B as appears in Table 8 is less than .09 by Fisher's exact test. As a consequence, even samples of a few hundred subjects may be informative. If a stability optimum does not exist, then stability can be at least appreciably improved by lengthening the test. If a stability optimum does exist, then other ways of possibly improving temporal stability should be tried.

PREDICTIVE VALIDITY

Once a test has been constructed, it may be used to predict performance on numerous external criteria. At this point the issue is no longer test construction (test length) but test scoring. The usual practice is to average all trials given. The rationale underlying this practice is the Spearman-Brown formula. By including all trials one maximizes reliability and stability and, hence, predictive validity for all criteria. We have already seen, however, that the assumptions of the S-B formula are systematically violated in performance testing. Furthermore, there is now a respectable body of literature to the effect that the differential content of a task changes with practice or, in the psychometric context, that the predictive validity of a performance test may vary from early to middle to late trials (Fleishman & Hempel, 1954; Ackerman, 1987).

Forward averages, of course, include only some trials, specifically, the first i, and they too may be correlated with an external criterion. When they are, the correlations (predictive validities) always rise at first and sometimes reach an optimum, after which they decrease. If, however, a

forward optimum in predictive validity exists, then averaging only those trials up to and including the optimum will yield a higher predictive validity than the usual practice. Since the differential composition of a test may change with practice and an external criterion may be most strongly related to those components of a test that predominate at the beginning (say) or in the middle of a practice series, stability and validity optima do not necessarily fall on the same trial. For the same reasons, the optimal forward average for purposes of prediction may vary from one external criterion to another.

Averaging from the first trial forward is only one way to generate a series of averages from a series of test trials. Another way is to average from the last trial backwards. Backward averages may also be correlated with an external criterion. When they are, the correlation (predictive validity) rises at first and may reach an optimum prior to the first trial. In these cases, as in the corresponding cases involving forward optima, averaging only those trials up to and including the optimum (following it in the practice series) yields a higher predictive validity than averaging all trials given. Backward optima are especially helpful in improving a test's validity when a forward validity optimum also exists.

Four of the ten Project-A tests have high conventional validities when performance four months later on the first retest session of Anti-Aircraft is used as the criterion. The two best predictors are Target Tracking 1 and 2, with validities of .696 in Sample A (both tests), .707 and .654 in Sample B. Second best are the two aiming tests (Cannon Shoot and Target Shoot), with validities of .594 and .510 in Sample A, .474 and .458 in Sample B. Serial averaging (forward or backward) yields very small and nonsignificant

improvements for these four tests. This result may not be happenstance. It could be that high correlations are less likely to change with practice (or test administration) than low ones.

Validities for the remaining six tests range from .333 (Memory) and .372 (Choice Reaction) down, in Samples A and B respectively. Five of the six tests, all but Choice Reaction, show forward optima in Sample A. The results for three of these tests (Simple Reaction, Number Memory, and Target ID) are presented in Figure 8. In all three cases validity follows a similar course, starting low, rising to an optimum, and then trailing off. In Target ID, for example, averaging the first five trials only yields the best result, .306. As more and more trials are added to the average, validity falls away until, when all 36 trials are averaged, it has fallen to .196.

Three of the six tests (Choice Reaction, Memory, and Number Memory) show backward optima in Sample A. The two memory tests show both a forward and a backward optimum. Number Memory, for example, has a forward optimum at Trial 16 and a backward optimum at Trial 4. Two optima, however, are one too many. Forward and backward averaging are only two out of a great many possible ways of searching out validity optima. Altogether there are 2^n or, in the case of Number Memory, 2^{28} possible combinations of trials. Were we to search all of these combinations, capitalization on chance would be extreme and shrinkage at cross-validation would also be extreme.

The most straightforward way to avoid these extremes is to limit the number of series that one examines. Accordingly, in defining an optimal validity average I have adopted the following three-step algorithm:

1, If neither a forward nor a backward optimum exists, then the optimal average is the average of all trials given (the conventional average).

2, If a forward optimum exists but not a backward optimum, the optimal average is the average of all trials from the first up to and including the optimal trial. Similarly, if a backward optimum exists but not a forward optimum, the optimal average is the average of all trials from the last back to the optimal trial.

3, If both forward and backward optima exist, the average of all trials spanned by the two optima is usually more valid than either the forward or backward optimum. If so, the optimal average is the spanning average. If not, the optimal average is the more valid of the forward and backward optima.

Implicit in this algorithm is a restriction to consecutive trials. In itself this restriction reduces the total number of possibilities to be searched from 2^n to $n(n+1)/2$. Capitalization on chance is still involved, of course, but its extent has been severely curtailed. Figure 9 illustrates the application of the algorithm to Number Memory. The two upward-pointing arrows mark the backward and forward optima, on the left and on the right respectively.

Table 9 presents validity information for the six tests other than the two tracking and the two aiming tests in Sample A. It includes: number of trials, the optimal average as reached by the algorithm just described, the validity of that average, the validity of the conventional average, and the difference (Δ) between the optimal and conventional averages. The final two columns contain the z-score (unit normal deviate) for Δ (a difference

between two correlations sharing a common variable and based on the same subjects) and the associated one-tailed significance level (Steiger, 1980, p. 247, Equation 14). Three of the six tests are significant at the .05 level and two others at the .10 level.

These results are for single tests. We may also ask how much difference optimal averaging makes in the validity of best composites of the Project-A tests. The validity of the six tests in Table 9 are representative of real-world, job-performance validities (Ghiselli, 1966; Schmidt, Hunter, & Pearlman, 1981). If these six tests are scored in the usual way, they yield a multiple correlation of .413. If the same six tests are scored by optimal averages, the multiple correlation is .496. The difference between the two multiple correlations yields a z-score of 2.03, significant at the .03 level. In short, for tests with representative validities optimal scoring may improve the validity of a test or battery by as much as .10. In practical terms, gains of this magnitude in tests designed to be used on a mass basis for personnel assignment are important. Since, moreover, these gains can be had "for nothing," there is no reason not to take them.

There remains, of course, the problem of capitalization on chance. Validity optima, unlike those for reliability and temporal stability, are not based on the entire set of results. A forward optimum, for example, is simply a forward average with a validity greater than that of any other forward average, specifically including the last, that is, the average of all trials administered. An average that meets this description could easily do so on the basis of a chance upward deflection. The algorithm for selecting a single optimum to some extend compounds this problem; and forming a multiple

composite compounds it further. It is necessary, therefore, always to check a validity optimum in an independent sample of subjects.

Table 10 presents such a check. The results are for Sample B, where the optimal averages are those obtained in Sample A (see Table 9) and the best composites, both for conventional and optimal scoring, are formed using the same weights as were found to maximize validity in Sample A. In three of the six tests the difference Δ favors conventional scoring and in the remaining three optimal scoring. The latter three differences, however, are all larger than the three differences that went "the wrong way" and one of them, that for Number Memory, is marginally significant, $p<.08$. The Δ for best composites goes in the right direction but by an amount 45% that in Sample A.

The directions and magnitudes of these differences are about what one would expect and probably representative of the gains to be had by optimal validity averaging. In absolute terms, however, gains ranging from .04 to .09 are not large. Significance cannot be expected unless sample size and, therefore, the power of the test are greater than they are in this study.

SUBSET ANALYSIS

The gains to be made by optimal scoring, though worthwhile, are not large. In order substantially to improve the validity of a test its content must be changed. In a few tests, Simple Reaction is a case in point, all trials are the same. In such a case one can always revise the test but one cannot distinguish any subset of trials in the existing test that has more validity than other subsets. In most tests, however, it is possible to distinguish such subsets. On any given trial the Project-A Memory Test, as pointed out earlier, presents the subject with 1, 3, or 5 stimuli to be retained in short-term memory. Accordingly, one can separate the trials of

Memory into three corresponding subsets. Other tests involve different types of stimuli (letters, numbers, or other symbols), moving or stationary stimuli, different time delays, or other parametric variations, which can be used to separate the trials into subsets.

"Subset analysis" serves the same ends in performance-test theory as item analysis does in conventional psychometrics. Both kinds of analysis concern the selection of some materials for inclusion in a test and others for exclusion, either in original development or in subsequent revision. The difference is that item analysis focuses on individual items and subset analysis on subsets of trials.

Implicit in this defining difference are three others which require comment. First, in a conventional analysis items are distinguished from one another by their relation to the criterion. Typically, one includes items with high validity and excludes items with little or no validity. Oftentimes, however, one does not have enough items of the first sort and wishes to create more. But what do these items have in common with one another apart from high individual item validity? In order to write more such items one must have an answer to this question; one must know what kinds of items to write. Accordingly, one conducts factor or cluster analysis in an effort to characterize the valid items. Once that is done, it may be possible to write more valid items--but not always. The problem of accurately characterizing valid items in a conventional analysis and then creating more items like them is a difficult and frequently frustrating task. In subset analysis it is no problem at all, because the subsets are distinguished from the beginning by observable, easily noted features. If trials with 3 stimuli are more valid than those with 1 or 5, it is a simple

matter to draw more combinations of 3 stimuli from a previously prepared list of suitable stimuli.

The second major difference between item and subset analysis concerns capitalization on chance. In a conventional analysis the number of items considered for inclusion in a test is usually much larger than the number selected. Hence, capitalization on chance and consequent shrinkage at cross-validation are pronounced. In subset analysis the number of subsets is usually small, not more than a handful. Hence, capitalization on chance, while it exists, is much less of a problem.

The third difference relates to substantive theory. Item analysis is a bitterly empirical procedure. Each item is related separately and directly to the criterion. The only link between one item and another is a latent and, therefore, nonmanipulable factor, an abstract idea. In subset analysis each subset is distinguished by an observable and manipulable feature that may well be or have been the subject of experimental study. It is this feature that links subset analysis to cognitive science. The number of objects, for example, that can be retained in short-term memory has been studied extensively by experimental investigators (Miller, 1956). One result is that seven is about the limit of the normal range. Hence, sets of nine or even seven stimuli to be retained in working memory would not discriminate among most subjects; too few people would respond correctly. Hence, too, sets of seven or more stimuli are unlikely to be either temporally stable or predictively valid.

In Perceptual S & A the subject is asked to compare two strings of symbols and indicate whether or not they are the same. The strings may be two, five, or nine stimuli in length. Accordingly, the trials of Perceptual

S & A break down into three subsets of 12 trials each. Figure 10 presents validity results for these three subsets in Sample A. The plotted points are correlations between forward averages within subsets (ignoring trials in other subsets) at test and the first retest session on Anti-Aircraft. Table 11 presents results for both samples and for temporal stability as well as predictive validity. Each subset in Table 11 is represented by the average of all 12 trials in the subset.

In Sample A the simplest of the three subsets (Set 2) has the best validity and the most complex (Set 9) the poorest. This same ordering reappears in Sample B and with similar spacing. In Sample A the difference between Sets 2 and 9 yields a unit normal deviate of $z = 2.06$. A two-tail test is appropriate in this case because any ordering could have been accommodated. In addition, three subsets are involved. Applying the Bonferroni procedure, one obtains a significance level of .15. In Sample B, however, the directions of all differences are stipulated in advance. Hence, though the unit normal deviate in Sample B, 1.80, is less than that in Sample A, its significance is greater, $p<.04$.

Table 11 illustrates two points. The first concerns scoring for validity. The validities for Subset 2 in Table 11 are better than the corresponding optimal validities for Perceptual S & A in Tables 9 and 10. In this instance, rescoring by subset analysis would have yielded a better result than optimal averaging in the complete set of 36 trials. Even under heavy constraints to guard against excessive capitalization on chance, scoring for validity in performance testing is a three-step process: optimal averaging in the set of all trials, subset analysis, and optimal averaging within subsets.

The second point illustrated by Table 11 is the possibility of using a subset analysis to restructure a performance test. Since Sets 2 and 5 are more valid than Set 9, the idea suggests itself of restructuring the test so that it includes 18 trials of length two, 18 of length five, and none of length nine. Such a restructured test would be expected to have a validity with conventional scoring on the order of .28 instead of the present validity with conventional scoring of .11. This suggestion, I will argue, should be rejected.

Perceptual S & A is far from being the best predictor of Anti-Aircraft in the Project-A battery. The two tracking tests and the two shooting tests are all much more valid than it is. There may well be, however, other criteria, clerical criteria, perhaps, for which Perceptual S & A is the primary predictor. It would make sense to restructure Perceptual S & A to make it more valid for such a criterion because in that case one would be improving the validity of the battery as a whole for that criterion. It would not make sense in the present case. The gain for the battery as a whole in predicting Anti-Aircraft would be at best very small; and restructuring Perceptual S & A to predict Anti-Aircraft might easily weaken its validity for those criteria it currently predicts better than other tests in the battery. Any such restructuring would be almost literally "penny wise and pound foolish."

The results for temporal stability in Table 11 underscore this hazard. In Sample A temporal stability is best for Set 9 and worst for Set 2. If this result held generally, then by restructuring Perceptual S & A to make it more valid for Anti-Aircraft, we would by the same stroke be making it less stable and, therefore, probably less predictive of those criteria where it

matters most. As it happens, the stability ordering reverses itself in Sample B. On balance, stability probably does not vary much one way or the other among the three subsets. Ideally, however, a restructuring to improve validity should improve stability as well.

Cannon Shoot provides an illustration. In both samples Cannon Shoot predicts the Anti-Aircraft criterion about as well as it predicts itself over the same interval of time (four months). Predictive validity in the two samples is .594 and .474; temporal stability in the same two samples (A and B respectively) is .534 and .545. The difference favors validity by .06 in Sample A and stability by .07 in Sample B. The main reason, it would appear, that Cannon Shoot doesn't predict Anti-Aircraft better (as well, for example, as the two tracking tests predict it) is its relatively low stability. If the stability of Cannon Shoot could be improved, its validity for Anti-Aircraft would likely also improve. It is possible that restructuring Cannon Shoot to make it more stable and more predictive of Anti-Aircraft might make it less valid for some other criteria. The validity, however, of Cannon Shoot for these other criteria would almost certainly be less than it is for Anti-Aircraft and less, too, than the validity of some other tests for those same criteria. For the battery as a whole the gains from restructuring Cannon Shoot would outweigh the losses.

The trials of Cannon Shoot differ in various ways: the position of the cannon, the speed and direction of the target. It would be possible, therefore, to separate trials into subsets on the basis of these variations. There would, however, be more than a few such subsets and, as will be seen, the main difference among trials is not so much any one of these variations as it is the trial's overall difficulty. Accordingly, the trials of Cannon

Shoot were divided into three groups of 12 trials each on the basis of mean performance at test.

Mean performance is, of course, a random variable. Hence, which trials are the 12 most difficult and which the 12 easiest may vary from sample to sample. In samples of even moderate size, however, this variation is likely to be minor. In the present data 11 of the 12 easiest trials and 11 of the 12 most difficult trials are the same in Samples A and B; 10 of the 12 middle-difficulty trials are the same in the two samples. In large samples there should be no variation at all from sample to sample, although even then the distinction between grouping trials by difficulty level and grouping them by fixed parameter settings should not be lost. In any case, grouping trials by difficulty level preserves the essential character of a subset analysis, because the difficulty of a trial can be determined independently of stability or validity considerations. In addition, one can easily add new trials by reproducing the combinations of position, speed, and direction that are known to be hard or easy in the existing tests.

Figure 11 presents validity results for the three difficulty subsets in Sample A. Table 12 presents both validities and stabilities for the three subsets in Samples A and B. Each subset is represented by the average of all 12 trials in the subset. In both samples Subset Hard is most predictive and most stable, while Subset Easy is least predictive and least stable. The differences, moreover, are large, ranging from a minimum of .149 (stability in Sample A) to a maximum of .412 (stability in Sample B). The differences for validity in Sample A and stability in Sample B are significant at the .02 level or better (z values of 2.89 and 3.51) even if one uses a two-tailed test and makes the Bonferroni correction. The conclusion is clear: if Cannon

Shoot were restructured to consist exclusively of difficult trials, its temporal stability and predictive validity could both be improved.

When a reliability or stability optimum is projected at some point past the end of the administered series, one cannot be sure that lengthening the test to that point will have the desired effect until one actually does it. Scoring for validity is not conditioned on any change in the existing test and, therefore, requires no check other than cross-validation. When a test is restructured following subset analysis, the possibility of contextual effects becomes a major threat. Given the results in Table 12, it seems likely that Cannon Shoot restructured to consist of 36 difficult trials would be both more valid and more stable than the existing test; but it may not be so. One can only be sure after the fact.

COMMENT

The results presented in this paper constitute only the beginning of a performance-test theory. How should the ability to perform well on a given test be modeled? How is fairness with respect to race and sex to be understood or lack of fairness to be detected, how are fair tests to be constructed? How are the tests to be protected against unfair advantage obtained by deliberate practice on the same or similar tests? These and many other questions remain to be addressed. At this point I will develop only one point. It concerns the greater importance of stability than reliability optima.

Learning or skill acquistion is one among many processes that can produce a superdiagonal correlation pattern. Almost any series of measurements ordered in space or time will generate correlations tending toward superdiagonality (Jones, 1960). Suppose, then, that in a given test

the intertrial correlations, apart from differences in individual baselines, are mediated by factors that come and go, for example, response sets, fatigue, distractions, variations in concentration, and the like. The closer two trials are the more likely it is that they will be affected by the same transient factors and, hence, covary together.

Now consider the forward averages for a given individual. Some transient factors will improve the individual's performance and others will worsen it. As a result, the subject's average performance will come more and more to approximate his or her baseline level as the test lengthens; and the ratio of true-score to total variance (reliability) will increase monotonically. In short, there will be no reliability optimum. This discussion began, however, with the assumption that intertrial correlations fell into a superdiagonal pattern because of transient factors. If so, then reliability as calculated by the Spearman-Brown formula could reach an optimum, even though, as has just been seen, no reliability optimum exists.

The hypothesis of superdiagonality mediated by transient factors has other consequences, however. If the hypothesis were true, there would be no mean change from test to retest. Nor would there be any stability optimum. Covariances between test and retest do not involve transient factors and should in the absence of enduring changes with practice be flat, while the variances of forward averages within test or retest decrease monotonically as j increases. Finally, stability ought not to place a bound on the validity of any test.

All three of these consequences are false. Mean performance improves from test to retest in all tests; contrary to expectation, temporal stability reaches an optimum in five of the ten tests; and the validity of at least one

test, Cannon Shoot, appears clearly to be limited by its relatively low stability. The superdiagonal patterns among intertrial correlations at test and retest cannot, therefore, be explained as resulting from transient factors, at least not entirely. Such factors may, however, still play a role additional to learning or skill acquisition. That is, both learning and nonlearning (transient) factors may be involved.

The most direct way to handle this complication is to rely primarily on stability optima, which cannot be explained in terms of transient factors and must, therefore, have their origins in more enduring changes with practice. This conclusion is reinforced by a further consideration. Reliability is a theoretical quantity, defined in terms of true-score and error variance. As such it is open to all the vagaries of interpretation. Temporal stability, on the other hand, is obtained empirically. When, therefore, it rises to an optimum and then decreases, as it does in Choice Reaction (Sample B), the fact is observable.

At one level, serial averaging is a prosaic data-processing procedure. It is based, however, on a view of performance testing that departs fundamentally from current test theory. The gist of that departure is not to replace one theory with another but to hybridize test theory with the study of individual differences in skill acquisition and retention. Conventional test theory is purely structural; time has no place in it. The study of skill acquisition and retention, however, is processual; everything in it is embedded in time and is, therefore, temporally ordered. Many parts of the hybrid theory presented in this paper come from its processual component: for example, the treatment in terms of trials, the centrality of order, or the recognition of established regularities such as superdiagonal form. The

overall approach is open, moreover, to further imports from the study of skill acquisition. Where stability optima exist, a possible solution (in addition to expanding some subsets and eliminating others) might be to administer the test in two or more well-separated bouts, each containing many fewer trials. Distribution and transfer effects, reminiscence, many possible results from cognitive science, may ultimately find a place in a theory of performance testing hybridized to include performance as well as test phenomena.

TECHNICAL NOTES

1. Average intertrial covariance within a test or retest session ($\overline{cov}_k$) may be calculated either directly or indirectly. If all subjects get a score on all trials, the two ways of calculating $\overline{cov}_k$ come to the same thing. However, when the dependent variable is mean decision time on only those trials where a subject responds correctly (as it is in five of the ten Project-A tests), this condition is not met. When calculated indirectly, $\overline{cov}_k$ is obtained from the formula for the variance of a forward average, that is,

$$Var_{FA} = \frac{\sum (FA - \overline{FA})^2}{k^2} \Big/ N-1$$

$$= \left(\overline{var}_k + (k-1)\,\overline{cov}_k\right)/k$$

or

$$\overline{cov}_k = \left(k\,var_{FA} - \overline{var}_k\right)/k-1,$$

where $\overline{var}_k$ and $\overline{cov}_k$ represent the average trial variance and covariance up to trial k. The merit of the indirect calculation is that when $\overline{cov}_k$ is so obtained, the decomposition formula for temporal stability remains exact. To

be consistent, the direct calculation is also used in making reliability projections.

2. Backward optima are not informative about how changes in test length might affect reliability or temporal stability. A forward average of, say, 5 trials retains its meaning (refers to the same trials) regardless of how many trials are ultimately given. A backward average of 5 trials, however, refers to trials 6-10 if 10 trials are given and to trials 11-15 if a total of 15 trials is given. A backward average changes its meaning when the total number of trials changes. As a consequence, no conclusion regarding changes in test length can be drawn from a backward reliability or stability optimum.

REFERENCES

Ackerman, P.L. (1987). Individual differences in skill learning: an integration of psychometric and information processing perspectives. Psychological Bulletin, 102, 3-27.

Bittner, A.C., Jr., Carter, R.C., Krause, M., & Harbeson, M.M. (1983). Performance Evaluation Tests for Environmental Research (PETER): Moran and computer batteries. Aviation, Space, and Environmental Medicine, 54, 923-928.

Eaton, N.K., Hanser, L.M., & Shields, J. (1985). Validating selection tests for job performance. In J. Zeidner (Ed.), Human productivity enhancement. Vol. 2, Acquisition and development of personnel. New York: Praeger.

Englund, C.E., Reeves, D.L., Shinglebecker, C.A., Thorne, D.R., Wilson, K.P., & Hegge, F.W. (1987). Unified tri-service cognitive performance assessment battery (UTC-PAB): I. Design and specification of the battery (Report No. 87-10). San Diego, CA: Naval Health Research Center.

Fleishman, E.A. & Hempel, W.E., Jr. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 19, 239-252.

Ghiselli, E.E. (1966). The validity of occupational aptitude tests. New York: Wiley.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley & Sons.

Jones, M.B. (1962). Practice as a process of simplication. Psychological Review, 69, 274-294.

Jones, M.B. (1969). Differential processes in acquisition. In E.A. Bilodeau (Ed.), Principles of skill acquisition. New York: Academic Press, 1969.

Jones, M.B. (1960). Molar correlational analysis. Pensacola, FL: U.S. Naval School of Aviation Medicine (Monograph 4).

Kennedy, R.S., Baltzley, D.R., Wilkes, R.L., & Koontz, L.A. (1989). Psychology of computer use: IX. A menu of self-administered microcomputer-based neurotoxicology tests. Perceptual and Motor Skills, 68, 1255-1272.

Kennedy, R.S., Bittner, A.C., Jr., Carter, R.C., Krause, M., Harbeson, M.M., McCafferty, D.B., Pepper, R.L., & Wiker, S.F. (1981). Performance Evaluation Tests for Environmental Research (PETER): Collected papers (NBDL-80R008). New Orleans, LA: Naval Biodynamics Laboratory.

Kyllonen, P.C., & Christal, R.E. (1989). Cognitive modeling of learning abilities: a status report of LAMP. In R. Dillon & J.W. Pellegrino (Eds.), Testing: theoretical and applied issues. New York, NY: Freeman.

Kyllonen, P.C. (1985). Theory-based cognitive assessment (AFHRL-TP-85-30). Brooks Air Force Base, TX: Air Force Human Resources Laboratory.

Melton, A.W., Ed. (1947). Apparatus tests. Washington, DC: U.S. Government Printing Office (AAF Aviation Psychology Program Research Report No. 4).

Messick, S., & Jungblut, A. (1981). Time and method in coaching for the SAT. Psychological Bulletin, 89, 191-216.

Miller, G.A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review. 63:81-97.

Peterson, N.G., Ed. (1987). Development and field test of the trial battery for Project A (Technical Report 739). Alexandria, VA: U.S. Army Research Institute.

Peterson, N.G., Hough, L.M., Dunnette, M.D., Rosse, R.L., Houston, J.S., Toquam, J.L., & Wing, H. (1990). Project A: specification of the predictor domain and development of new selection/classification tests. Personnel Psychology 43, 247-276.

Reynolds, B. (1952). The effect of learning on the predictability of psychomotor performance. Journal of Experimental Psychology, 43, 341-348.

Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: a red herring. Journal of Applied Psychology 66, 166-185.

Steiger, J.H. (1980). Tests for comparing elements of a correlation matrix. Psychological Bulletin, 87, 245-251.

Wing, H. (1980). Practice effects with traditional test items. Applied Psychological Measurement, 4, 141-155.

Table 1. Number of trials and total length of time for the 10 Project-A, computer-administered tests.

Test	Number of Trials	Total Time (mins)
Simple Reaction Time	10	2
Choice Reaction Time	30	3
Memory Test	36	7
Target Tracking 1	18	8
Perceptual Speed & Accuracy	36	6
Target Tracking 2	18	7
Number Memory	28	10
Cannon Shoot	36	7
Target Identification	36	4
Target Shoot	30	5

Table 2. Comparison of Army and Hershey results with the Project-A computer-administered tests.[1]

Test/Measure	Army/ Hershey	$\bar{X}$	s	Reliability	Temporal Stability
Simple Reaction (mean dec. time)	Army[2]	31.84	14.82	.88	.23
	Hershey	29.38	4.94	.88	.50
Choice Reaction (mean dec. time)	Army	40.83	9.77	.97	.69
	Hershey	36.54	6.48	.97	.77
Memory (mean dec. time)	Army	87.72	24.03	.96	.66
	Hershey	70.98	17.43	.97	.69
Target Tracking 1 (mean ln dist. + 1)	Army	2.98	0.49	.98	.74
	Hershey	2.77	0.43	.98	.87
Perceptual S & A (mean dec. time)	Army	236.91	63.38	.94	.63
	Hershey	202.42	47.10	.95	.73
Target Tracking 2 (mean ln dist. + 1)	Army	3.70	0.51	.98	.85
	Hershey	3.45	0.52	.98	.91
Number Memory (final resp. time mean)	Army	160.70	42.63	.88	.62
	Hershey	118.39	27.89	.91	.69
Cannon Shoot (mean abs. time disc.)	Army	43.94	9.57	.65	.52
	Hershey	43.80	8.52	.51	.53
Target ID (mean dec. time)	Army	193.65	63.13	.97	.78
	Hershey	163.84	45.08	.95	.71
Target Shoot (mean ln dist. + 1)	Army	2.17	0.24	.74	.37
	Hershey	2.14	0.20	.71	.70

1 All times are in hundredths of a second. Logs are natural logs.

2 Simple Reaction in the Army battery has 15 trials. Number of trials in the remaining tests are the same in the Army as in the Hershey battery The Army retest results are based on overlapping samples of 468 to 487 subjects.

Table 3. Means and standard deviation on Memory at test and retest, broken down by subset and sample. Individual scores are averages of all 12 trials in a subset.

Sample	Session	Subset	s	$\bar{X}$	$s/\bar{X}$
A	Test	1	152.1	581.3	.261
		3	192.7	749.3	.257
		5	203.3	806.0	.252
	Retest	1	147.2	575.7	.256
		3	151.0	708.1	.213
		5	193.8	802.5	.241
B	Test	1	154.2	607.2	.254
		3	188.3	781.3	.241
		5	234.3	869.5	.269
	Retest	1	145.3	608.4	.239
		3	181.1	758.7	.239
		5	183.2	835.2	.219

Table 4. Intertrial correlations for Perceptual Speed and Accuracy in Sample A, test session, averaged in blocks of four trials.

	Trials								
Trials	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32	33-36
1-4	(.39)	.34	.32	.23	.21	.24	.18	.22	.17
5-8		(.54)	.40	.45	.34	.40	.32	.35	.31
9-12			(.43)	.42	.39	.42	.34	.39	.28
13-16				(.54)	.44	.38	.43	.41	.31
17-20					(.45)	.41	.41	.48	.34
21-24						(.41)	.38	.42	.30
25-28							(.43)	.45	.38
29-32								(.50)	.37
33-36									(.35)

Table 5. Hypothetical correlations among seven trials of practice, together with the average correlation ($\bar{r}_i$) and reliability (R_i) as calculated by the Spearman-Brown formula up to a given trial.

Trial	Trial 1	2	3	4	5	6	7
1	-	.80	.65	.50	.35	.20	.05
2		-	.80	.65	.50	.35	.20
3			-	.80	.65	.50	.35
4				-	.80	.65	.50
5					-	.80	.65
6						-	.80
7							-
$\bar{r}_i$	-	.80	.75	.70	.65	.60	.55
R_i	.800	.889	.900	.903	.902	.900	.895

Table 6. Reliability results for the ten Project-A computer-administered tests, using pooled averages ($\tilde{r}_i$) and weighted regression (Sample A).

Test	No. of Trials	$b \times 10^3$	i^*	R_{i^*} or r_1
Simple Reaction	10	- 14.42	14.8	.874
Choice Reaction	30	- 1.05	195.3	.988
Memory	36	- 1.16	159.4	.977
Target Tracking 1	18	- 2.46	100.3	.992
Perceptual S & A	36	+ 0.16	none	1.000
Target Tracking 2	18	+ 1.69	none	1.000
Number Memory	28	- 1.21	96.2	.921
Cannon Shoot	36	- 1.09	29.9	.510
Target ID	36	- 0.50	310.9	.987
Target Shoot	30	- 1.82	33.8	.704

Table 7. Stability results for the ten Project A computer-administered tests (Sample A).

Test	No. of Trials	Slope x 10^3 b	c	i*	R_i* or r_{11}
Simple Reaction	10	-17.2[1]	+7.5	24.3	.566
Choice Reaction	30	- 0.9	-0.4	45.7	.775
Memory	36	- 0.5	+4.1	none	1.000
Target Tracking 1	18	- 0.8	+0.7	none	1.000
Perceptual S & A	36	+ 1.7	+2.4	none	1.000
Target Tracking 2	18	- 0.2	+0.0	none	indet.[2]
Number Memory	28	- 2.8	-1.6	31.1	.670
Cannon Shoot	36	- 0.5	+1.1	43.8	.583
Target ID	36	- 0.0	+4.0	none	1.000
Target Shoot	30	+ 1.8	+2.9	48.7	.762

1 b_2 indicates slope for the pooled correlations at retest. For the corresponding slope at test, see Table 6. c indicates slope for the covariance ratio.

2 The test regression line reaches unity while stability is still rising. Hence, any "final" stability is speculative and is best put down as "indeterminate."

Table 8. Optimal trial numbers (i*) for reliability and stability in Samples A and B.

	Reliability		Stability	
Test	Sample A	Sample B	Sample A	Sample B
Simple Reaction	14.8	133.9	24.3	26.1
Choice Reaction	195.3	473.6	45.7	15.2
Memory	159.4	89.0	none	none
Target Tracking 1	100.3	81.1	none	none
Perceptual S & A	none	164.9	none	37.8
Target Tracking 2	none	none	indet.	none
Number Memory	96.2	none	31.1	19.1
Cannon Shoot	22.9	64.8	43.8	74.7
Target ID	310.9	304.1	none	60.4
Target Shoot	33.8	45.4	48.7	58.1

Table 9. Optimal averages for the Project-A tests in predicting performance in the first retest session on Anti-Aircraft (Sample A).

Test	No. of Trials	Optimal Average Start	Optimal Average End	Predictive Validity Opt.	Predictive Validity Conv.	Δ	z	p
Simple Reaction	10	1	6	.299	.251	.048	1.47	<.08
Choice Reaction	30	28	30	.162	.117	.045	0.84	n.s.
Memory	36	6	19	.291	.237	.054	2.06	<.02
Perceptual S & A	36	1	9	.222	.107	.115	1.52	<.07
Number Memory	28	4	16	.406	.333	.073	2.05	<.03
Target ID	36	1	5	.306	.196	.110	1.94	<.03

Table 10. Cross-validation results for the six tests with optimal averages and their best linear composite (Sample B).

Test	Composite		
	Opt.	Conv.	Δ
Simple Reaction	.211	.251	-.040
Choice Reaction	.424	.372	+.052
Memory	.347	.349	-.002
Perceptual S & A	.230	.174	+.056
Number Memory	.296	.210	+.086
Target ID	.298	.314	-.016
Best Composite	.393	.356	+.037

Table 11. Subset analysis for Perceptual S & A in two independent samples of 102 subjects each. Each subset is represented by the average of all 12 trials in the subset.

Result	Sample A	Sample B
Predictive validity:		
Subset 2	.231	.267
Subset 5	.177	.225
Subset 9	.055	.101
Temporal stability:		
Subset 2	.590	.694
Subset 5	.633	.612
Subset 9	.724	.609

Table 12. Subset analysis for Cannon Shoot in two independent samples of 102 subjects each. Each subset is represented by the average of all 12 trials in the subset.

	Sample	
Result	A	B
Predictive validity:		
Easy 12	.263	.243
Average 12	.331	.334
Hard 12	.565	.395
Temporal stability:		
Easy 12	.295	.160
Average 12	.398	.323
Hard 12	.444	.572

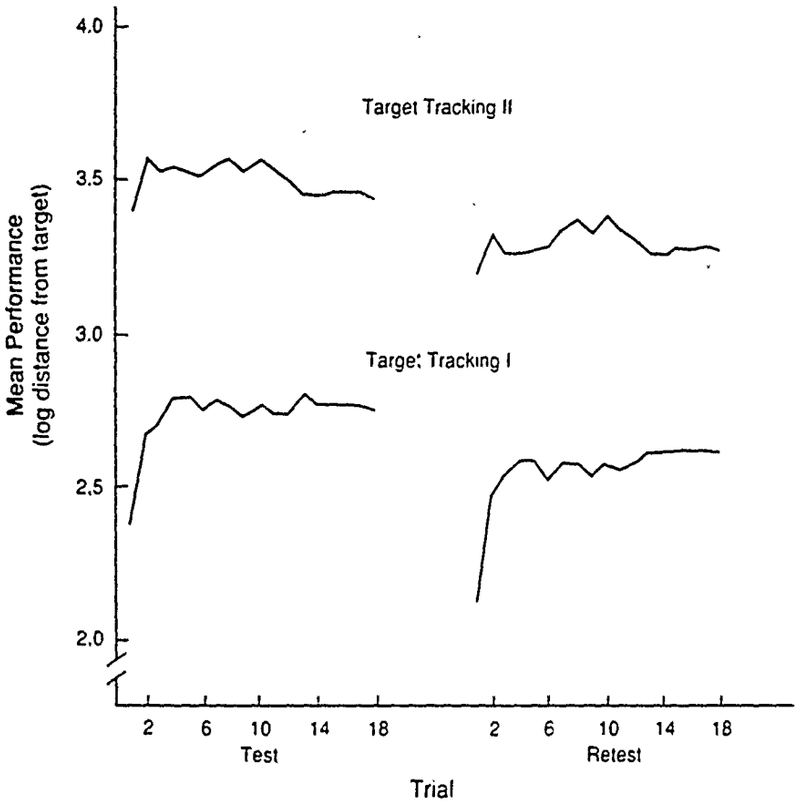

Figure 1. Mean performance at test and retest on the two tracking tests (Sample A).

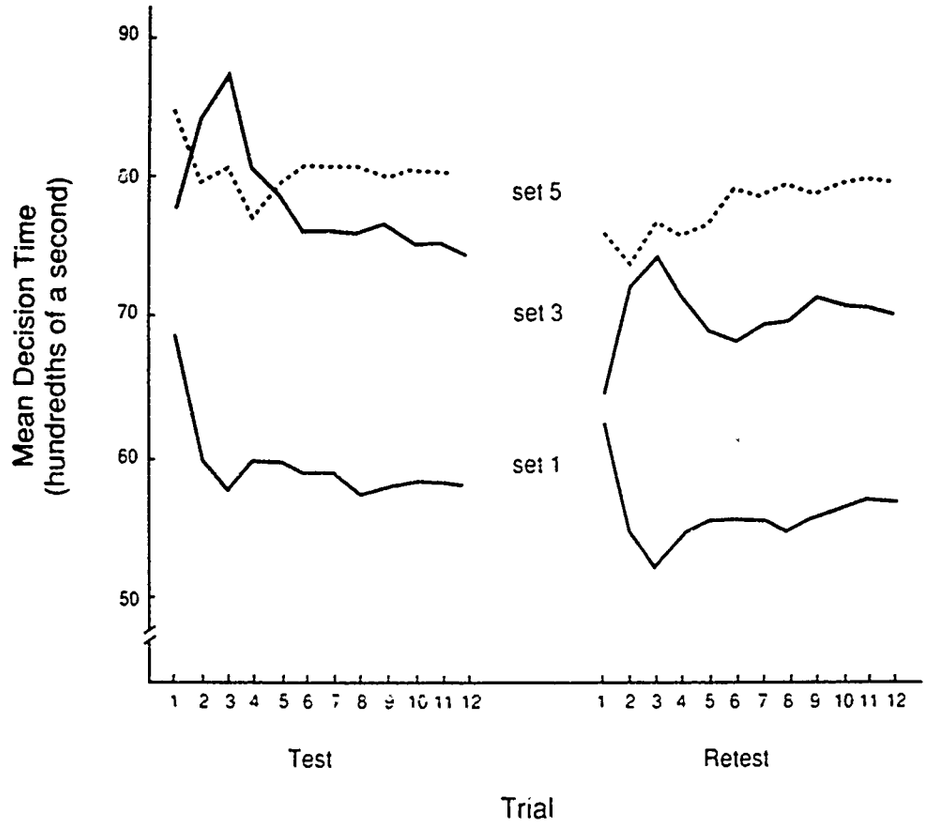

Figure 2. Mean performance at test and retest on Memory, by subset (Sample A).

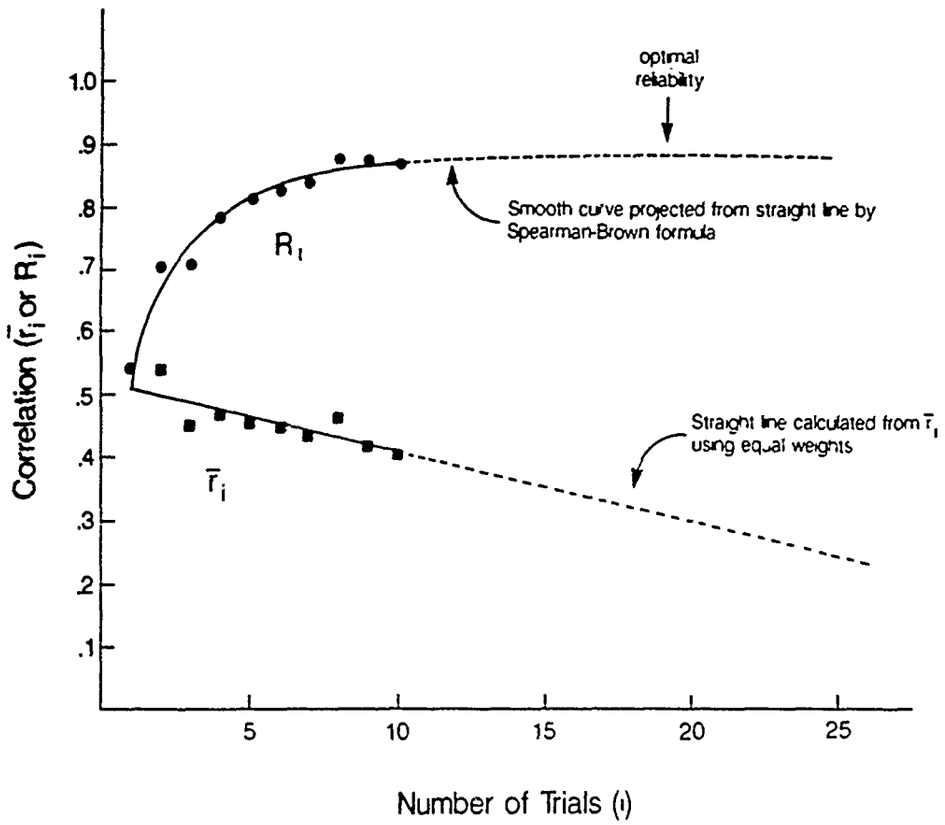

Figure 3. Average correlation and Spearman-Brown reliability up to trial i for Simple Reaction in Sample A. The straight line has been calculated weighting all average correlations equally.

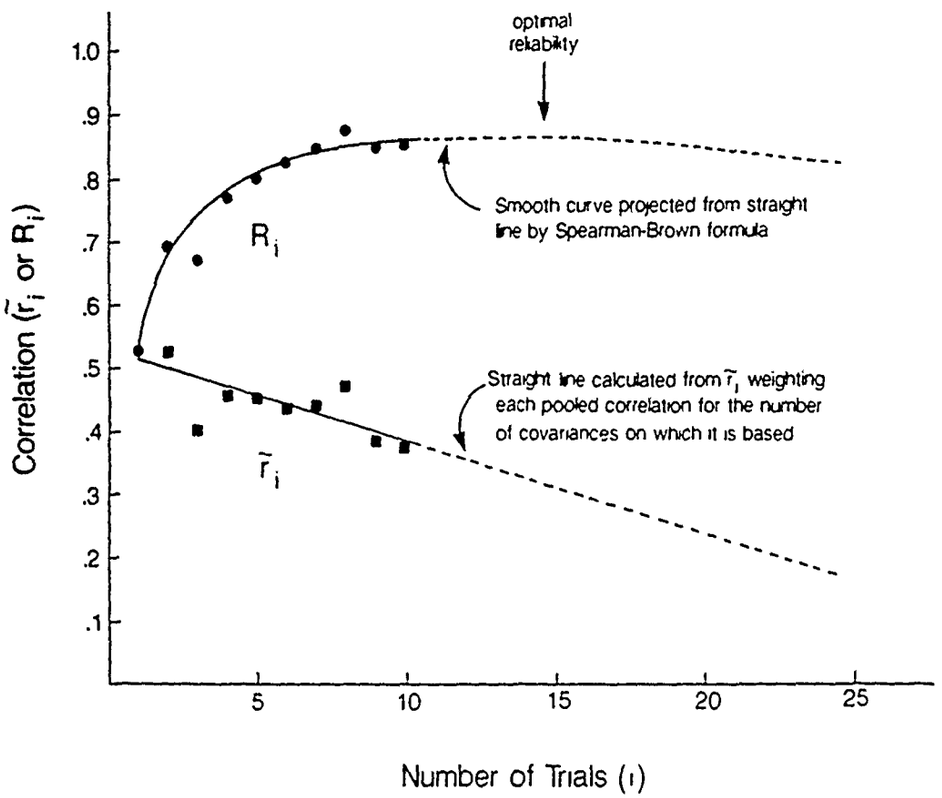

Figure 4. Pooled correlations ($\overline{cov}/\overline{var}$) and Spearman-Brown reliability up to trial $\underline{i}$ for Simple Reaction in Sample A. The straight line has been calculated weighting each pooled correlation for the number of covariances on which it is based.

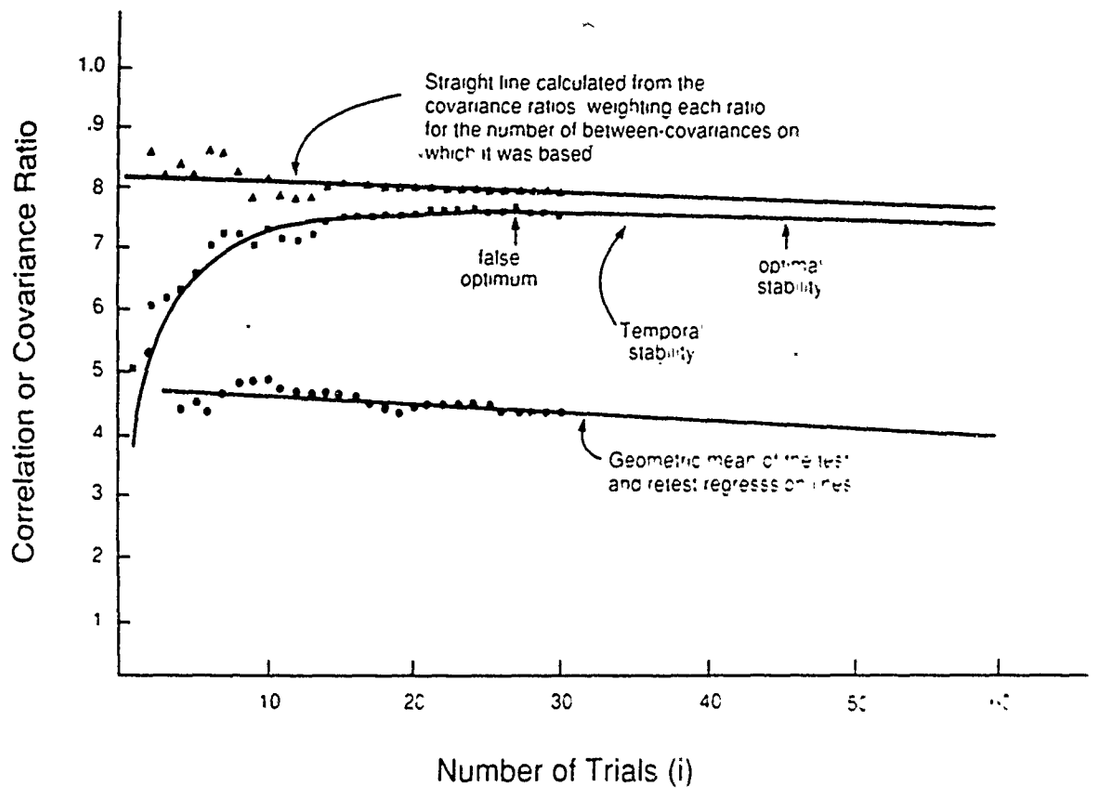

Figure 5. Temporal-stability results for Choice Reaction in Sample A.

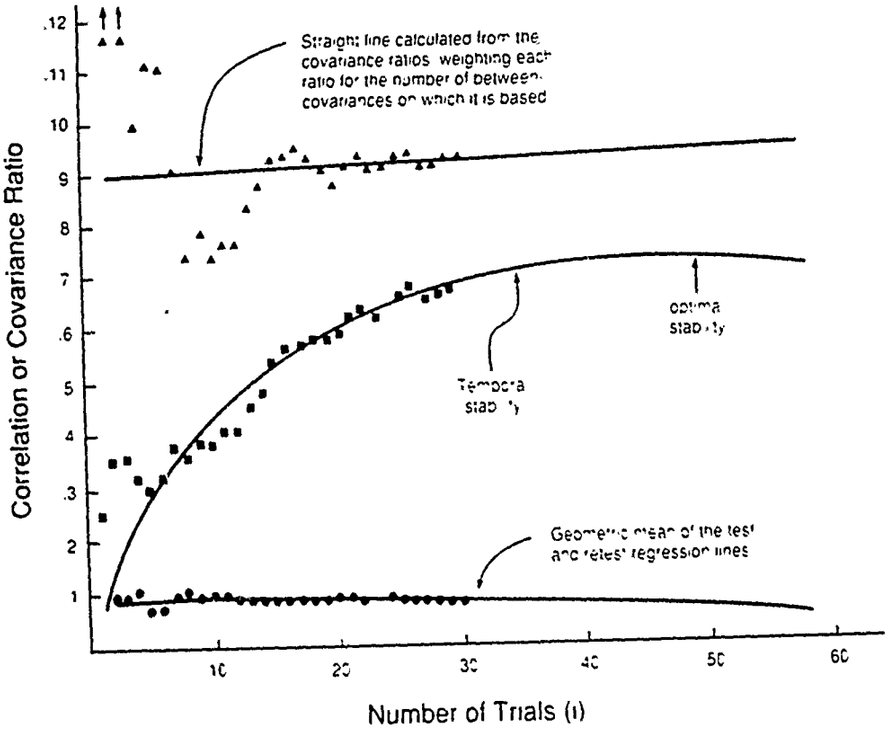

Figure 6. Temporal-stability results for Target Shoot in Sample A.

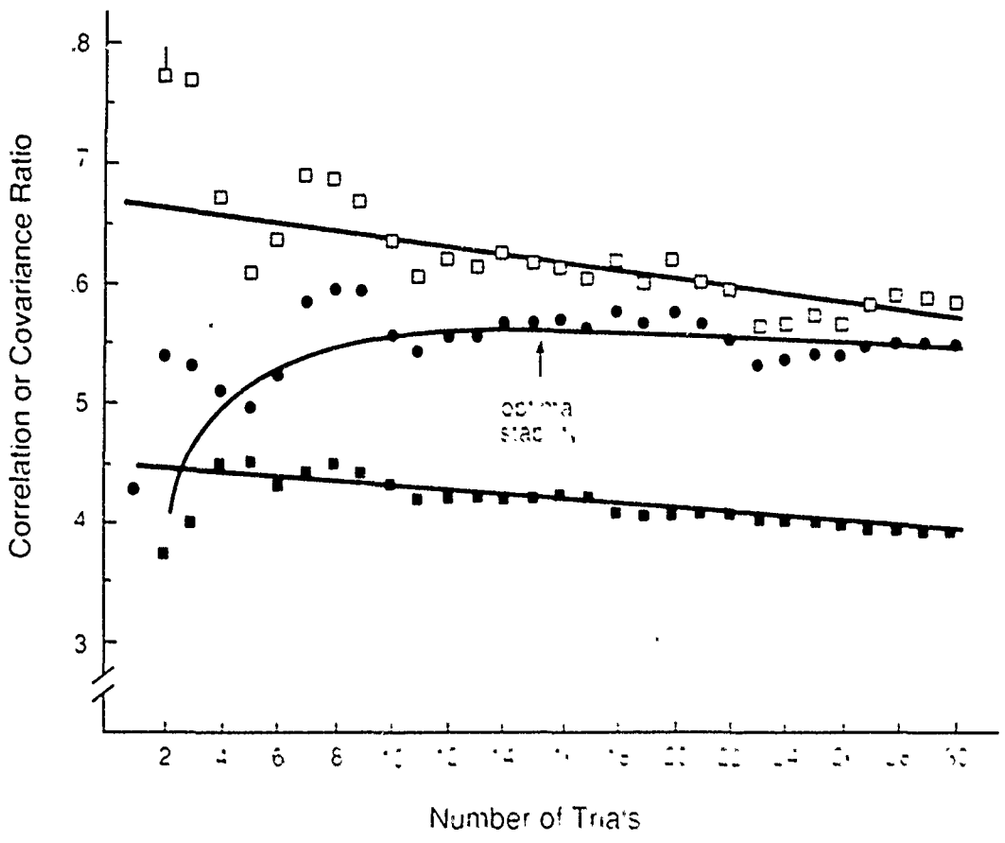

Figure 7. Temporal-stability results for Choice Reaction in Sample B.

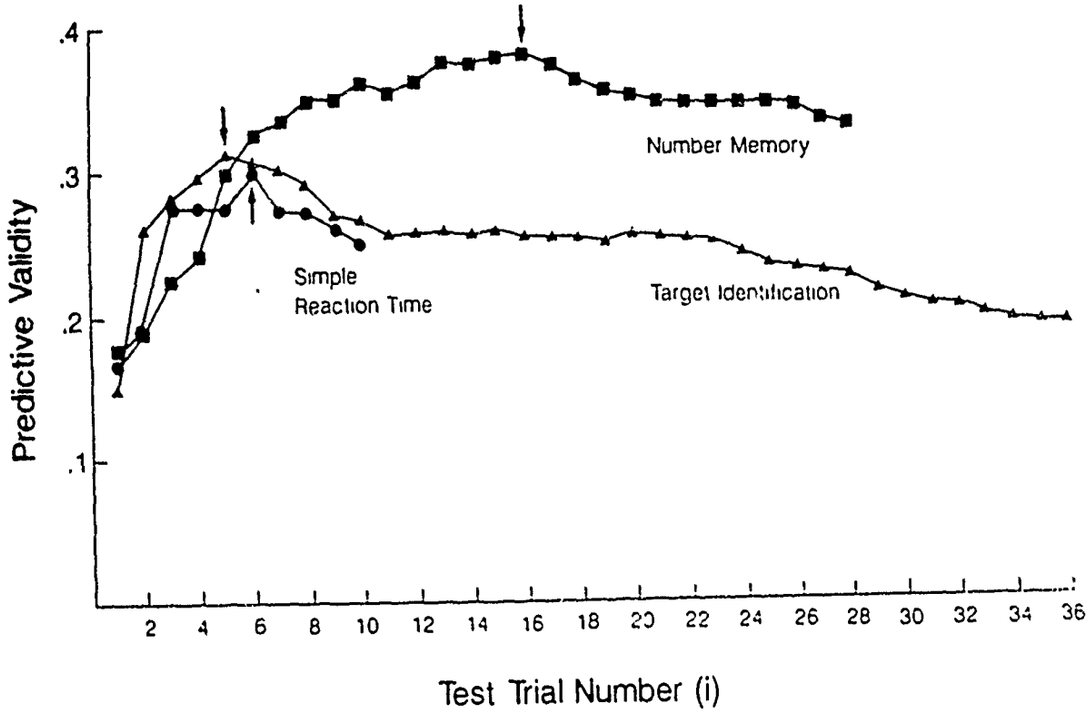

Figure 8. Forward validity results for Simple Reaction, Number Memory, and Target Identification in Sample A.

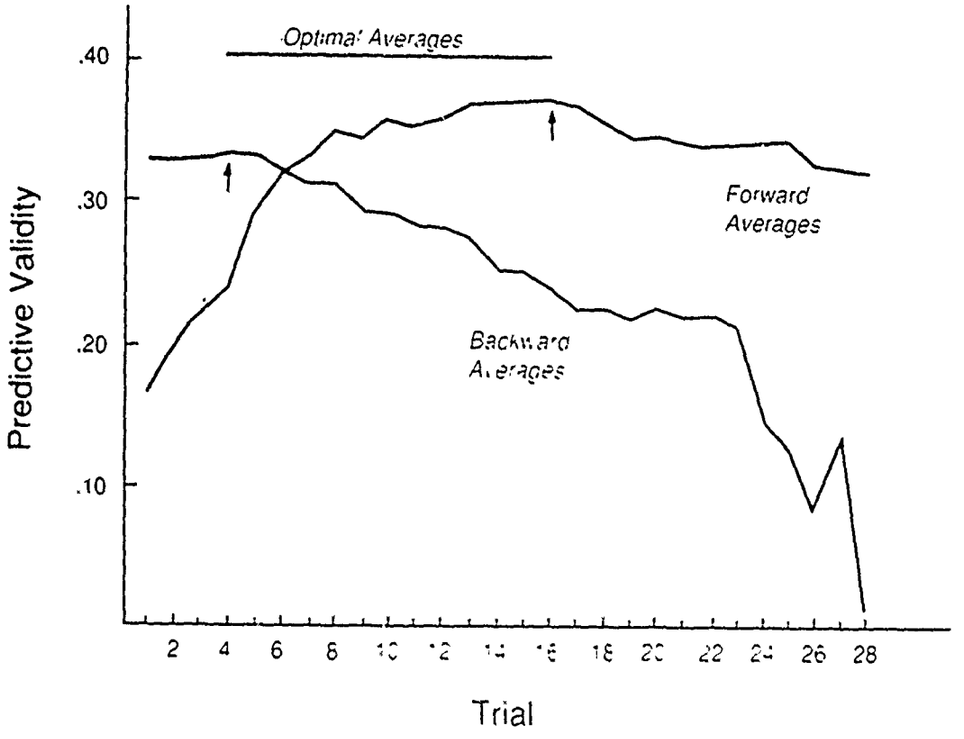

Figure 9. Forward, backward, and optimal validities for Number Memory in Sample A.

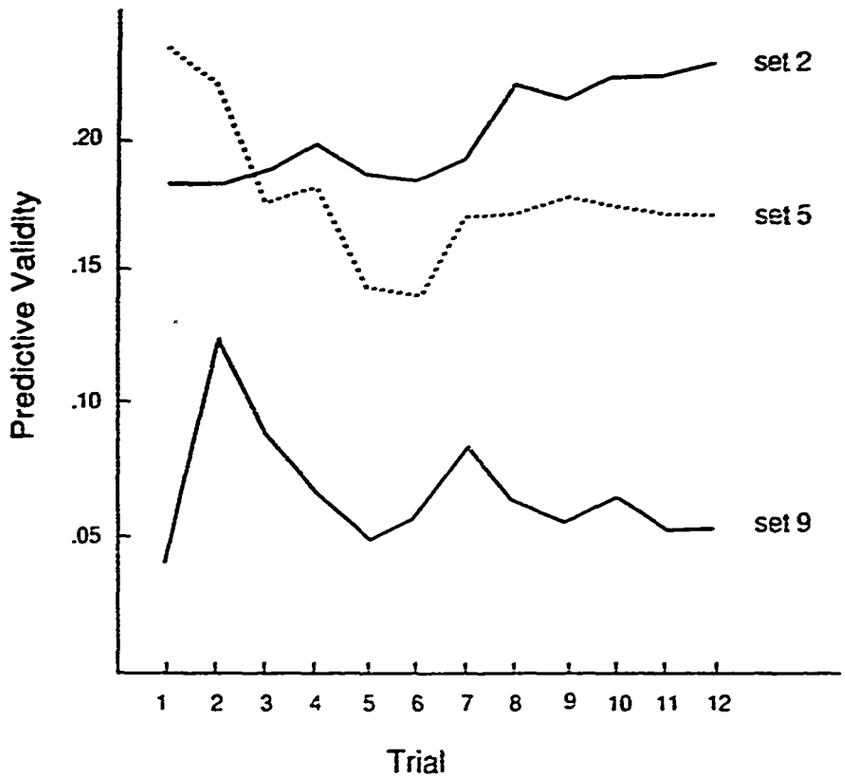

Figure 10. Validity results for Perceptual Speed and Accuracy in Sample A, by subset.

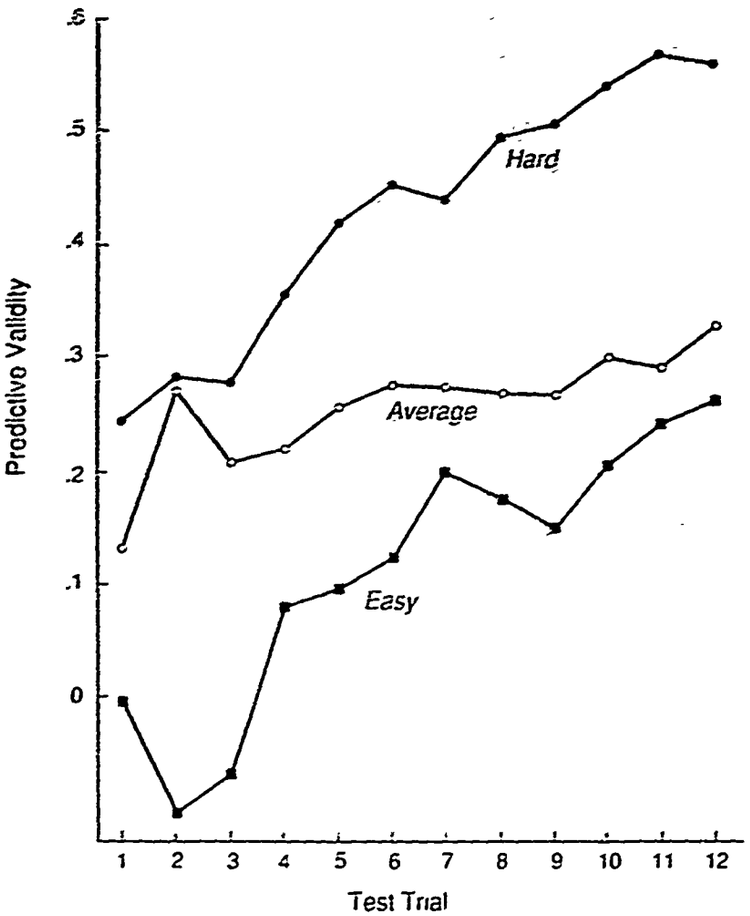

Figure 11. Validity results for Cannon Shoot in Sample A, by subset.